VIA ECF

June 13, 2025

Hon. Ona T. Wang
Southern District of New York

RE: *In re OpenAI Copyright Litigation*, 25-md-3143 (SHS) (OTW)
This Document Relates To: 23-CV-11195

Dear Magistrate Judge Wang:

Pursuant to this Court's June 11, 2025 Order (Dkt. 161), The New York Times Company ("The Times"), Daily News, LP, et al. ("Daily News Plaintiffs"), and The Center for Investigative Reporting ("CIR") (collectively the "News Plaintiffs") submit their response to OpenAI's sampling proposal set forth in OpenAI's June 6, 2025, Proposed Order Regarding Conversation Data Sampling. Dkt. 104 (the "Proposed Order"). As discussed in News Plaintiffs' letter dated June 6, 2025, OpenAI did not provide News Plaintiffs a copy of OpenAI's proposal in advance of OpenAI's submission to this Court. Dkt. 98.

News Plaintiffs set forth below specific comments about the scope and statistical reliability of OpenAI's sampling proposal, together with proposed modifications to OpenAI's Proposed Order attached hereto as Exhibit A (marked up) and Exhibit B (clean).

I. OpenAI's Proposed "Test" and "Control" Samples Are Not Effective for Identifying Relevant Content

OpenAI's proposed Test and Control samples are not sufficient because they: (1) exclude relevant and searchable data fields in OpenAI's "Conversation Data" logs, and (2) do not consider variances between and within distinct data populations.

With reference to Section 1.a of the Proposed Order, the definition of Conversation Data should be clarified to include both prompts and output pairs for a given conversation, along with all of the logs, records, and data fields ordinarily collected in the individual rows of consumer output log data. This data includes, for example, the intermediate queries and responses resulting from retrieval augmented generation ("RAG") functionality, including any URLs and excerpts referenced in the RAG response, along with metadata associated with each conversation such as classifiers.

News Plaintiffs have two concerns regarding OpenAI's proposed samples in Section 1.b of the Proposed Order.

First, the Test Sample and Control Sample include overlapping populations of data because the Test Sample includes "all" Conversation Data, whereas the Control Sample includes Conversation Data that is ordinarily collected.[1] Having overlapping data populations reduces the power of the sampling exercise and the variances within the two Samples. For example, if someone wanted to evaluate whether men and women have different heights, OpenAI is proposing a sampling exercise

[1] Conversation Data that is "ordinarily collected" refers to the Conversation Data not marked for deletion. In other words, this is the Conversation Data that OpenAI would not have deleted, irrespective of the Court's Preservation Order.

where the Test Sample includes men and women, and the Control Sample includes only men. With respect to data marked for deletion, this issue is further compounded because there are two potentially different populations: (1) Conversation Data generated using the "Temporary Chat" feature; and (2) Conversation Data marked for user-initiated deletion.

Second, OpenAI's proposal does not take into account variances between the Control Sample over time due to the block lists and other post-litigation measures OpenAI put in place to suppress the regurgitation of News Plaintiffs' content and to make it more difficult to elicit verbatim output of News Plaintiffs' content.

To address these sampling issues, News Plaintiffs propose four sample populations (including one that pre-dates all of the MDL actions filed against OpenAI) as follows:

i. Test Sample 1 - the sample population for Conversation Data generated through the "Temporary Chat" feature from April 14, 2025 through May 14, 2025.

ii. Test Sample 2 – the sample population for Conversation Data marked for user-initiated deletion from April 14, 2025 through May 14, 2025.

iii. Control Sample 1 – the sample population for Conversation Data from April 14, 2025 through May 14, 2025 that: (i) was not generated through the "Temporary Chat" feature and (ii) is not subject to a user-initiated deletion.

iv. Control Sample 2 – the sample population for retained Conversation Data from April 14, 2023 through May 14, 2023.

II. The Sampling Methodology from the *Anthropic* Case is Likely Not Relevant to the Sampling Exercise for the Output Log Data Marked for Deletion

News Plaintiffs do not have sufficient information to ascertain whether OpenAI's proposed sample size in Section 1.c of the Proposed Order will result in a statistically valid sample size for the four sample populations identified above. For example, News Plaintiffs do not know the volume of the different populations of data under consideration or the prevalence rates of infringement and news-related use cases in and among those data populations. But solely for purposes of expeditiously conducting the sampling exercise for the output log data marked for deletion, and without prejudice to News Plaintiffs' right to request a different sampling methodology for the retained Conversation Data for other aspects of discovery, News Plaintiffs accept OpenAI's proposal of five (5) million rows of Conversation Data for each of the Test Sample 1, Test Sample 2, Control Sample 1, and Control Sample 2, which amounts to twenty (20) million rows of Conversation Data across the four samples.

The court in *Concord Music Grp., Inc. v. Anthropic PBC*, Dkt. 377, Case No. 5:24-cv-03811 (N.D. Cal. May 23, 2025) ("*Anthropic*") adopted both parties' proposal to use Cochran's formula to calculate a statistically significant sample size from a population of six months' worth of Anthropic's output log data. Cochran's formula includes three variables: (i) a "Z-score" corresponding to a desired confidence level, (ii) an estimated prevalence of the characteristic to be studied in the sample population (e.g., the prevalence of infringement), and (iii) a margin of error corresponding to an acceptable sampling error. The parties agreed to a Z-score of 1.96 corresponding to a 95% confidence level and an expected prevalence of 0.00006 (the prevalence of end-users searching for song lyrics), but disputed whether the margin of error should fall in the 5-25% range (plaintiffs' expert) or a 10-50% range (Anthropic's expert). To resolve the dispute, the court ordered a margin of error of

approximately 11.3%, which results in a sample size of 5 million. News Plaintiffs agree that Cochran's formula is a helpful tool to arrive at a statistically significant sample size for certain data populations, and used Cochran's formula to formulate their proposal to OpenAI for sampling historical ChatGPT Free, Pro, and Plus that OpenAI has retained. Dkt. 64-2 at pp. 2-3 (News Plaintiffs' May 20, 2025 Letter to OpenAI). OpenAI has not responded to that proposal.

III. OpenAI's Search Strategy is Too Narrow

News Plaintiffs have three concerns regarding OpenAI's search methodology set forth in Section 2 of the Proposed Order.

First, OpenAI's proposal to run keyword searches over **just the prompts** in the Conversation Data excludes instances where an output or RAG response hits on a keyword even where the prompt does not expressly call out one of the keywords. For example, a user may ask about current events without specifically referencing one of the News Plaintiffs, and receive a response that includes or was based on News Plaintiffs' intellectual property. (*E.g.*, a request for "What happened between Trump and Musk yesterday?" elicits a verbatim output from, or summary of, a News Plaintiffs' article.) Accordingly, OpenAI should run keyword searches over all of the data fields (including the outputs and RAG responses)[2]—not just prompts—in the Conversation Data for the four proposed samples.

Second, OpenAI's proposal does not capture broader news-related uses of ChatGPT that may not hit on any of the enumerated keywords but still implicate News Plaintiffs' intellectual property rights. For example, a user may ask ChatGPT to generate a news article in the style of a famous reporter, or to summarize or generate a derivative work of a copyrighted news article. To evaluate such uses, OpenAI should provide a breakdown of hits on news-related use cases, such as by searching the Samples for news-related classifiers or other metadata fields that indicate news-related use cases. OpenAI appears to employ such classifiers for at least some of its output log data and News Plaintiffs have proposed using these classifiers to identify relevant output log data. *See* Ex. C at 3 (email correspondence).

Third, OpenAI's proposed keywords are too narrow. News Plaintiffs propose additional keyword search terms enumerated in the Appendix to Exhibits A and B, which include additional variations on (a) the names of News Plaintiffs' publications, and (b) the News Plaintiffs' domains in OpenAI's proposal, as well as the following additional terms: (c) "news," "journalism," and "magazines" (d) the websites and blocklists OpenAI has identified, and (e) a list of known "pink-slime" journalism sites that plagiarize News Plaintiffs' works.

***

News Plaintiffs are prepared to address OpenAI's Proposed Order at the upcoming June 25, 2025 discovery conference or at another time convenient for the Court.

[2] To the extent included in the rows of Conversation Data, OpenAI may exclude from the search fields containing block lists or other filters that reference News Plaintiffs' names or websites from the search.

June 13, 2025

Respectfully submitted,

/s/ Steven Lieberman
Steven Lieberman
Rothwell, Figg, Ernst & Manbeck, P.C.

/s/ Ian Crosby
Ian B. Crosby
Susman Godfrey L.L.P.

/s/ Matt Topic
Matt Topic
Loevy & Loevy

cc: All Counsel of Record (via ECF)